Inférence Bayésienne, probabilités conditionnelles

Table des Matières

Inférence Bayésienne, probabilités conditionnelles

I. Tests de dépistage

I. A. Efficacité d'un dépistage

☞**Activité 1**

Les tests rapides d'orientation diagnostique (TROD) sont des tests qui permettent la détection d'anticorps et/ou d'antigènes (par exemple détection des anticorps VIH 1 et 2 avec ou sans antigène). Ils sont techniquement très faciles à utiliser de façon individuelle et donnent un résultat très rapide lisible à l'oeil nu. Une étude sur un TROD du VIH, le test DetermineTM HIV-1/2, a été réalisée au laboratoire de virologie de l'hôpital Saint-Antoine (Paris) entre le 1^er juin 2002 et le 31 janvier 2010 sur 1 429 personnes. Les résultats sont consignés dans le tableau ci-dessous.

	Patients infectés	Patients non infectés	Total
TROD positif	39	8	47
TROD négatif	1	1381	1382
Total	40	1389	1429

Source : Immuno-analyse et Biologie spécialisée, Volume 26, Février 2011, p-23-26

On choisit une fiche patient au hasard, chaque fiche a la même probabilité d'être choisie, il y a **équiprobabilité** des fiches patients.

1. (a) Quelle est la probabilité d'obtenir la fiche d'un patient ***faux positif*** ?
   (b) Quelle est la probabilité d'obtenir la fiche d'un patient ***faux négatif*** ?
2. (a) Sachant que la fiche est celle d'un patient infecté, quelle est la probabilité que la fiche soit celle d'un patient vrai positif ?
   On appelle cette probabilité ***la sensibilité notée Se****.*
   (b) Sachant que la fiche est celle d'un patient non infecté, quelle est la probabilité qu'une fiche soit celle d'un faux négatif ?
   On appelle cette probabilité ***la spécificité notée Sp****.*
   (c) Pour un test quelconque, vers quelle valeur doivent tendre **Se** et **Sp** pour que le test puisse être considéré fiable ?

   À titre d'exemple, l'Organisation Mondiale de la Santé (OMS) recommande d'utiliser des tests de dépistage du VIH de sensibilité supérieure à 98% et de spécificité supérieure à 99%
3. (a) Sachant que la fiche est celle d'un patient positif, quelle est la probabilité que la fiche soit celle d'un patient infecté ?
   On appelle cette probabilité ***la valeur prédictive positive notée VPP***
   (b) Sachant que la fiche est celle d'un patient négatif, quelle est la probabilité que la fiche soit celle d'un patient non infecté ?
   On appelle cette probabilité ***la valeur prédictive négative notée VPN***
   (c) Quel commentaire pouvez-vous faire sur les deux valeurs obtenues VPP et VPN ?

I. B. Influence de la prévalence

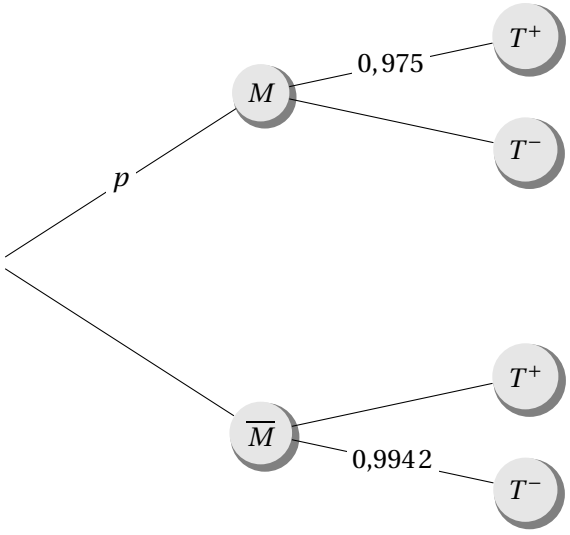
☞Activité 2

La prévalence p est la proportion de personnes infectées dans la population considérée. Dans l'activité précédente : $p = \dfrac{40}{1\,429} \simeq 2,8\%$.

Que se passe-t-il si on utilise le même test (même sensibilité et même spécificité) dans une population où la prévalence est différente ?

On représente la situation à l'aide de l'arbre suivant où :

- M désigne l'événement « le patient est infecté »,
- T^+ « le résultat du test est positif »,
- T^- « le résultat du test est négatif »,



1. Compléter les pondérations de l'arbre pondéré de probabilités.
2. (a) Que représente l'événement $\overline{M} \cap T^+$? Déterminer une expression de la probabilité de cet événement en fonction de p.
   (b) Que représente l'événement $M \cap T^-$? Déterminer une expression de la probabilité de cet événement en fonction de p.
3. (a) Déterminer une expression de VPP en fonction de p. On note f la fonction associée, elle est définie et dérivable sur $[\,0;\,1]$.
   (b) Déterminer une expression de VPN en fonction de p.On note g la fonction associée, elle est définie et dérivable sur $[\,0;\,1]$.
4. Étudier les variations des fonctions f et g, commenter.
   Pour information, la prévalence du VIH en France est de $0,26\%$
5. *Pour aller plus loin* : pour quelle valeur de p, $VPP > 0,5$? Justifier.

II. Spam

☞**Activité 3**

Qu'est-ce qu'un spam ?

Pour la petite histoire, SPAM est le nom d'une marque de jambon épicé (SPiced hAM en anglais) et son utilisation pour désigner les courriels indésirables a pour origine un sketch des Monty Python : un couple s'installe à la table d'une taverne, demande le menu et se voit harceler par la tenancière qui ne cesse de leur proposer du SPAM.

C'est le nom qu'ont pris les courriels indésirables dans nos boîtes de messagerie électronique, que l'on appelle aussi des pourriels. La plupart des boîtes de messagerie sont équipées d'un système de filtrage qui permet de les repérer et, par exemple, de les ranger directement dans un dossier dédié.

Le premier programme de filtrage du courrier électronique utilisant l'inférence bayésienne fut le programme iFile de Jason RENNIE en 1996. Le principe du filtrage bayésien s'est alors développé et des variantes de la technique de base ont donné naissance à de nombreux produits et logiciels « anti-spam ».

Principe du filtrage :

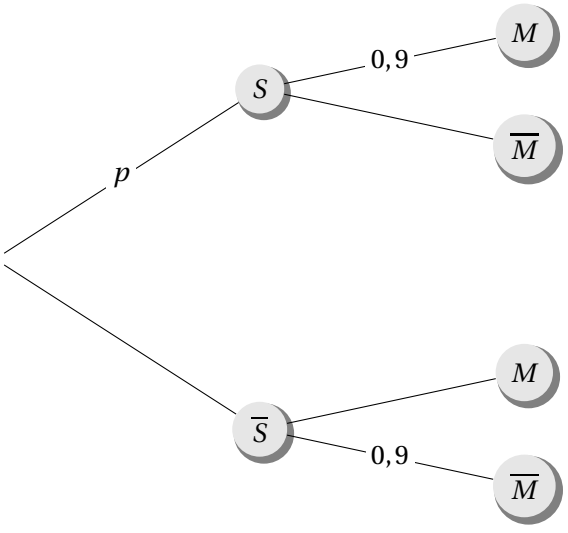
Le principe de base du filtrage bayésien consiste d'abord à repérer les mots plus fréquents dans les spams puis à évaluer en fonction des mots contenus dans un message si celui-ci doit être ou non considérer comme un spam. Ce procédé nécessite donc l'intervention de l'utilisateur de la messagerie électronique. Ce dernier doit en effet indiquer les messages qu'il considère être des spams. Cela permet au fur et à mesure d'évaluer la probabilité que certains mots apparaissent dans un spam. C'est donc un procédé évolutif qui peut s'adapter à chaque utilisateur.

Exemple de modèle :

En simplifiant le filtrage sur un mot donné, on peut modéliser la situation de la façon suivante. On choisit un message au hasard dans une boîte de messagerie et on considère les événements suivants :

- S : « le message est un spam »,
- M : « le message contient le mot donné ».

On représente cette situation avec l'arbre suivant :



1. Pour $p = 0,5$,
   (a) Calculer $P(M)$
   (b) En déduire la spamicité du mot, à savoir $P_M(S)$.
2. On admet que $p \in [0,55\ ;\ 0,95]$.
   (a) Exprimer la spamicité du mot en fonction de p. On note f cette fonction définie et dérivable sur l'intervalle $[0,55\ ;\ 0,95]$
   (b) Étudier les variations de la fonction f. Commenter.

III. Formule de Bayes

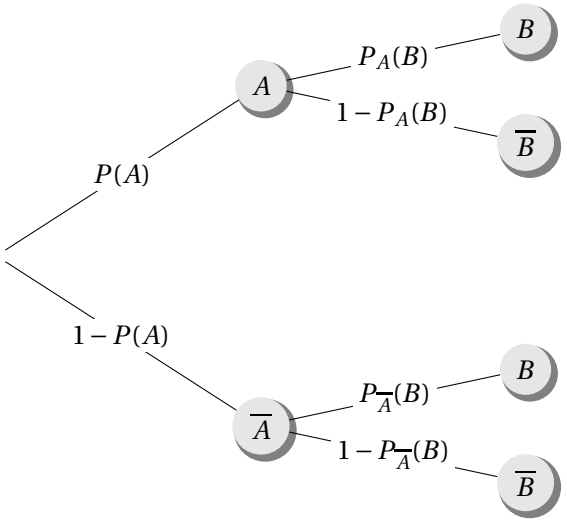
III. A. Histoire

La formule et la théorie de l'inférence bayésienne doivent leur nom au révérend Thomas Bayes, mathématicien et pasteur britannique né à Londres aux environs de l'année 1702 et mort en 1761. Il rédige un Essai sur la manière de résoudre un problème dans la théorie des risques (Essay Towards Solving a Problem in the Doctrine of Chances)dans lequel il applique la formule de Bayes, cet Essai sera publié à titre posthume en 1763. En réalité, la formule de Bayes était déjà connue d'autres mathématiciens comme Bernoulli et De Moivre. Bayes ne l'avait utilisée que dans un cas particulier sans en voir toute la généralité. Celle-ci sera explicitée par Pierre-Simon de Laplace (1749-1827) en 1774 et en donnera toute la portée. À l'époque, il s'agit alors plutôt de l'appliquer dans des problèmes liés aux jeux de hasard. Aujourd'hui l'inférence bayésienne a de nombreuses applications en médecine, en informatique mais aussi en sciences cognitives et dans les théories de l'apprentissage comme le deep learning.

III. B. La formule de Bayes

Théorème

Soit deux événements A et B d'un univers Ω.
On connait *à priori* $P(A)$, $P_A(B)$, $P_{\overline{A}}(B)$.



La probabilité *à postériori* $P_B(A)$, appelée **formule de Bayes** est :

$$P_B(A) = \frac{P(A) \times P_A(B)}{P(A) \times P_A(B) + P(\overline{A}) \times P_{\overline{A}}(B)}$$

Démonstration 1

Laissée en exercice

Exemple

Avec les notations de la définition et les notations de l'activité 2 :
Si l'on sait qu'une maladie touche 1% de la population, on *à priori* 1% de chance de la contracter ($P(A) = P(M) = 0,01$). Si on réalise un test de dépistage de cette maladie et qu'il est positif, la probabilité qu'on soit atteint *à posteriori*, $P_A(B) = P_{T^+}(M) = VPP$ c'est-à-dire une fois le résultat du test connu, devient plus grande, sans pour autant atteindre 100%, les tests n'étant jamais fiables à 100%.